Big Data and Computing Visions

www.bidacv.com

Big. Data. Comp. Vis. Vol. 4, No. 1 (2024) 12–21.

Paper Type: Original Article

Machine Learning and AI for Advancing Parkinson's Disease Diagnosis: Exploring Promising Applications

Shohreh Abdollahi[1], Ramin Safa[1,*]

[1] Department of Computer Engineering, Ayandegan Institute of Higher Education, Tonekabon, Iran; shohreabdollahi@yahoo.com; safa@aihe.ac.ir.

Citation:

Received: 19 November 2023 *Revised: 20 Januray 2024* *Accepted: 03 Februray 2024*	Abdollahi, S., & Safa, R. (2024). Machine learning and AI for advancing Parkinson's disease diagnosis: exploring promising applications. *Big data and computing visions*, *4*(1), 12-21.

Abstract

Parkinson's disease is a progressive neurodegenerative disorder affecting millions worldwide. It is characterized by tremors, stiffness, and movement problems, significantly impacting the quality of life for individuals affected by it. Early detection of this condition is crucial for effective management and treatment. Machine learning algorithms have shown promise in various medical applications, including disease detection. These algorithms can analyze large datasets, extract relevant patterns and features, and make accurate predictions. In the context of this neurological disorder, machine learning techniques offer the potential to develop efficient and reliable diagnostic tools. This study investigates the efficacy of three widely employed algorithms – Logistic Regression, Support Vector Machine, and Artificial Neural Network – in detecting Parkinson's disease using speech-related features. The analysis reveals that Artificial Neural Network achieves the highest accuracy of 92.4%, surpassing Logistic Regression and Support Vector Machine. Accordingly, further research should explore deep learning methods and integrate additional data sources, such as gait analysis and genetic markers, to enhance diagnostic capabilities for Parkinson's disease.

Keywords: Parkinson's disease, Machine learning, Classification, Comparative analysis.

1 | Introduction

Parkinson's disease is a debilitating neurodegenerative condition that affects the motor system, leading to symptoms such as tremors, stiffness, and movement problems [1]. It is a progressive disorder that significantly impacts the quality of life for individuals affected by it. Early detection of Parkinson's disease is crucial as it enables timely intervention and tailored treatment plans, improving patient outcomes [2]. In recent years, machine learning algorithms have shown promise in various medical applications, including disease detection.

Corresponding Author: safa@aihe.ac.ir

https://doi.org/10.22105/bdcv.2024.458879.1180

These algorithms can analyze large datasets, extract relevant patterns and features, and make accurate predictions [3]. In the context of Parkinson's disease, machine learning techniques offer the potential to develop efficient and reliable diagnostic tools [4].

The primary objective of this study is to expedite the quick and efficient detection of Parkinson's disease by utilizing machine learning approaches. Specifically, we investigate the efficacy of three widely employed algorithms. Logistic regression [5], Support Vector Machine (SVM) [6], and Artificial Neural Network (ANN) [7]. By meticulously comparing and contrasting these algorithms' performance, we aim to identify the most effective and accurate algorithm for the early detection of Parkinson's disease. This analysis will contribute valuable insights into enhancing the diagnostic process and improving patient outcomes. Parkinson's disease presents unique challenges for detection, as speech impairments in affected individuals often characterize it. These speech impairments result from an altered frequency spectrum and reduced control of vocal production due to the loss of limb control. Therefore, we focus on speech-related features as inputs to our machine-learning algorithms. The analysis of these features can provide valuable insights and discriminatory information for distinguishing between individuals with and without Parkinson's disease.

To accomplish our study objective, we utilize a dataset consisting of speech recordings from individuals diagnosed with Parkinson's disease, along with a control group of healthy individuals. We extract a range of acoustic features, such as vocal tremor, pitch variation, and other relevant parameters, from these recordings. These features serve as inputs to our machine-learning algorithms. By training and evaluating the Logistic Regression, SVM, and ANN algorithms on the Parkinson's dataset, we aim to develop a robust and accurate classification model. Performance evaluation metrics, including accuracy, precision, recall, and F1 score, will be employed to assess the effectiveness of the algorithms.

The organization of this paper is as follows. Section 2 provides an overview of Parkinson's disease, its symptoms, and the importance of early detection. Section 3 describes the preprocessing steps and the methodology. Section 4 presents the selection and training phase, where we compare logistic regression, SVM, and ANN algorithms. Evaluation metrics and performance analysis are discussed in Section 5. Finally, Section 6 concludes the paper and outlines future directions for research in Parkinson's disease detection using machine learning algorithms.

2 | Background and Related Work

Machine learning and AI are increasingly recognized as transformative tools in the healthcare industry. Their importance lies in their ability to analyze vast patient data and extract valuable insights. By leveraging machine learning algorithms, healthcare professionals can make more accurate diagnoses, develop personalized treatment plans, and improve patient outcomes. Integrating machine learning and AI in healthcare brings numerous benefits [8]. These technologies can automate repetitive tasks, such as administrative work and data entry, freeing healthcare professionals to focus on providing quality care. Machine learning algorithms can also analyze complex datasets, including medical images and genomic data, to identify patterns and detect early signs of diseases. This early detection can lead to timely interventions and improved patient outcomes.

Machine learning and AI enable precision medicine by considering individual patient characteristics and tailoring treatments accordingly. These technologies can predict treatment responses and optimize therapy plans by analyzing patient-specific data, such as genetic information and medical history. This personalized approach enhances treatment efficacy and minimizes adverse effects [9]. Furthermore, Machine learning and AI have the potential to revolutionize healthcare research and public health management. They can analyze large-scale health data to identify trends, risk factors, and epidemiological patterns. This information aids in formulating public health policies, resource allocation, and disease prevention strategies. Machine learning algorithms also support clinical trials by identifying suitable patient cohorts, predicting treatment outcomes, and optimizing trial designs [7]. Parkinson's disease is one of the fields where AI can significantly contribute.

Parkinson's disease is a complex neurodegenerative disorder that affects millions of people worldwide. It is characterized by the progressive degeneration of dopaminergic neurons in the substantia nigra region of the brain, leading to a deficiency of dopamine, a neurotransmitter responsible for motor control [3]. This deficiency results in the motor symptoms associated with Parkinson's disease, such as tremors, bradykinesia (slowness of movement), rigidity, and postural instability. Early detection of Parkinson's disease is essential for effective management and treatment. Traditionally, diagnosis has relied on clinical assessment by neurologists, which involves evaluating the patient's medical history, conducting physical examinations, and assessing motor symptoms [3]. However, this approach can be subjective and may not always provide accurate results.

In recent years, machine learning techniques have emerged as promising tools for Parkinson's disease detection. These algorithms have the potential to analyze large amounts of data and identify patterns that may not be apparent to human observers. Several studies have explored using machine learning algorithms for Parkinson's disease detection, focusing on various data modalities such as speech, gait, and handwriting. For instance, researchers have investigated the use of speech analysis for Parkinson's disease detection. Speech impairments, such as altered frequency spectrum and reduced vocal control, are common in individuals with Parkinson's disease [10]. By extracting acoustic features from speech recordings and applying machine learning algorithms, researchers have achieved promising results in differentiating between individuals with Parkinson's disease and healthy controls.

Gait analysis is another modality explored for Parkinson's disease detection. Parkinson's disease can cause changes in gait patterns, including reduced stride length, shuffling gait, and asymmetrical movement [11]. By collecting gait data using wearable sensors and applying machine learning algorithms, researchers have developed models that can distinguish between individuals with Parkinson's disease and those without. Furthermore, studies have explored the use of handwriting analysis as a potential marker for Parkinson's disease. Handwriting characteristics, such as tremors and micrographia (reduced letter size), can indicate Parkinson's disease [4]. Machine learning algorithms applied to handwriting data have shown promise in accurately identifying individuals with Parkinson's disease [12].

In this study, we aim to compare the performance of three commonly used machine learning algorithms for disease detection and provide a roadmap for analyzing these datasets. We begin by extracting speech-related features from speech recordings, which serve as inputs to our classification model. The extracted features are then utilized to train and evaluate the performance of the aforementioned algorithms. By executing and comparing the results obtained from these algorithms, we can assess their effectiveness in accurately detecting Parkinson's disease. Through this comprehensive approach, we aim to contribute to advancing Parkinson's disease detection using machine learning techniques and provide insights into the most suitable algorithm for this task.

3 | Methodology

In this paper, our main objective is to identify whether a person is affected by Parkinson's disease. To achieve this, we have reviewed a pre-existing study that utilized deep belief networks, support vector regression, linear regression, and neuro-fuzzy techniques for disease detection. However, for our research, we have focused on three different algorithms: logistic regression, SVM, and ANN. We will begin by training these algorithms using a dataset from the UC Irvine machine learning repository[1], which includes various features and parameters related to patient health, such as voice analysis, speech testing, blood reports, and X-ray reports. As illustrated in *Fig. 1*, these data will serve as inputs to our classification models.

[1] https://archive.ics.uci.edu/dataset/174/parkinsons

Logistic regression is a probabilistic algorithm that assesses the likelihood of a specific event occurring in a given scenario. In our study, we will utilize logistic regression to predict the probability of a person being prone to Parkinson's disease. By analyzing the characteristic events and parameters associated with the disease, we can determine the regression model's output, indicating whether the prediction is positive (yes) or negative (no). The SVM is a robust algorithm capable of solving various machine learning problems, particularly regression tasks, through classification processes.

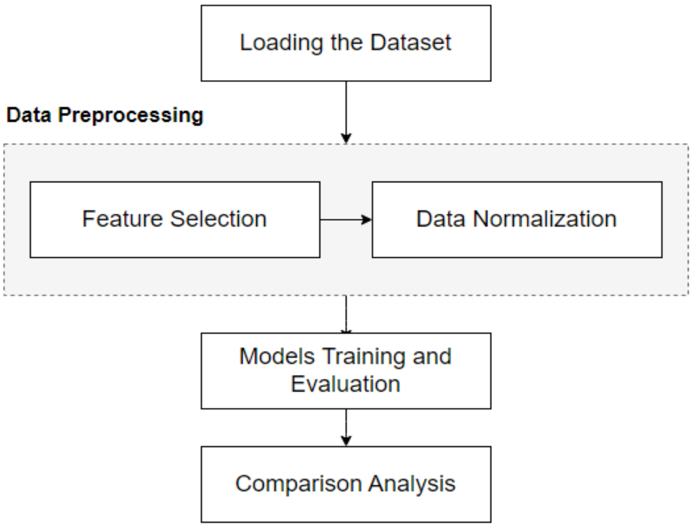

Fig. 1. Conceptual flow diagram of the overall process.

Our research will use SVM to classify patients based on their health-related symptoms, such as voice analysis, speech testing, and other medical reports. By comparing the symptoms of affected patients with those of healthy individuals, we will determine the relevance of the symptoms to Parkinson's disease and calculate the percentage likelihood of a person having the disease based on these classifications. ANNs are highly versatile algorithms that mimic the functioning of the human brain. They are beneficial for solving complex patterns and issues. In our study, we will leverage ANN, inspired by the human brain's structure and functionality, to analyze patient data, including health records and reports. By processing these inputs through the neural network, we can accurately determine the probability of a person being affected by Parkinson's disease. Compared to the other algorithms, ANN offers higher accuracy in disease detection. To evaluate the performance of the three selected algorithms, we will compare their results using a confusion matrix and test data. The confusion matrix will allow us to assess each algorithm's accuracy, precision, recall, and F1 score in detecting Parkinson's disease. By analyzing these metrics, we will identify the algorithm that demonstrates the highest accuracy in disease prediction and recommend its use for accurate disease detection. Our methodology involves training logistic regression, SVM, and ANN algorithms using patient data, evaluating their performance using appropriate metrics, and selecting the most accurate algorithm for Parkinson's disease detection.

4 | Experimental Analysis

For this study, the related dataset of patients with Parkinson's disease can be found in the following sources, including medical institutions specializing in Parkinson's disease research and treatment. These sources may include reputable hospitals, clinics, or research centers that have collected and curated datasets comprising various health-related parameters such as voice analysis, speech testing results, blood reports, and X-ray reports. The specific details regarding the availability and access to these datasets can be obtained by contacting the respective institutions or referring to their published research studies and data repositories.

Ensuring that the dataset collection follows ethical considerations and privacy protections to safeguard patient confidentiality is crucial. A visualization of the features of the data used in this research is provided in *Fig. 2*.

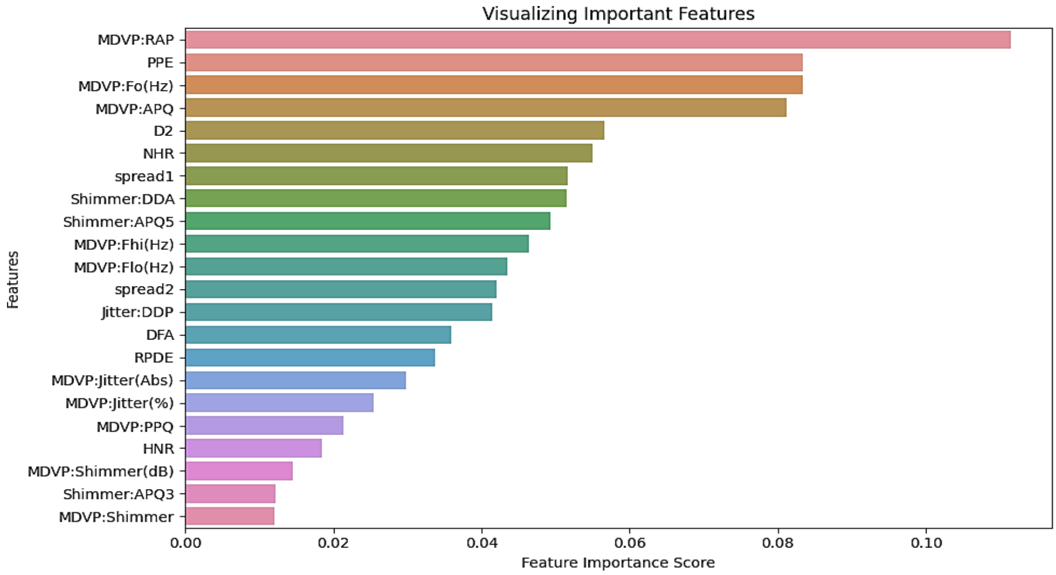

Fig 2. Visualization of important features.

Before training the algorithms, the dataset underwent several preprocessing steps to ensure data quality and compatibility [13]. Missing values within the dataset were addressed using imputation techniques like mean or median imputation, where the missing values were replaced with the mean or median value of the respective feature. Categorical variables, on the other hand, were encoded using suitable methods such as one-hot encoding or label encoding, depending on the variable's nature and the algorithm's requirements. This step transformed categorical variables into numerical representations that the algorithms can effectively utilize. Furthermore, feature scaling was applied to normalize the numerical features, ensuring they were on a similar scale and preventing any particular feature from dominating the learning process due to its magnitude. Overall, these preprocessing steps were crucial in preparing the dataset for the subsequent training and analysis stages, optimizing the performance and reliability of the algorithms utilized.

Algorithm 1 outlines the steps for preprocessing a dataset before training machine learning models. It includes handling missing values, encoding categorical variables, and feature scaling.

Algorithm 1. Data preprocessing steps.

```
Handling Missing Values:
    For each feature in the dataset:
        Calculate the number of missing values.
        If missing values exist:
            If the feature is numerical:
                Impute missing values with the mean of the feature.
            If the feature is categorical:
                Impute missing values with the mode of the feature.
Encoding Categorical Variables:
    For each categorical feature in the dataset:
        If the number of categories is 2:
            Perform label encoding on the feature.
        If the number of categories is more than 2:
            Perform one-hot encoding on the feature.
Feature Scaling:
    For each numerical feature in the dataset:
        Apply feature scaling to normalize the values and ensure they are on a similar scale.
```

4.1 | Evaluation Metrics

Several evaluation metrics were employed to evaluate the performance of the trained models. These metrics included accuracy, precision, recall, and F1-score. The accuracy represents the overall correctness of the predictions. It is calculated as the ratio of the number of correct predictions (true positives and true negatives) to the total number of predictions.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN), \tag{1}$$

where:

TP: true positives (correctly predicted positives).

TN: true negatives (correctly predicted negatives).

FP: false positives (incorrectly predicted positives).

FN: false negatives (incorrectly predicted negatives).

Precision measures the proportion of true positives among the predicted positives. It is calculated as the ratio of true positives to the sum of true positives and false positives.

$$\text{Precision} = TP / (TP + FP). \tag{2}$$

Recall, also known as sensitivity or true positive rate, calculates the proportion of true positives identified among the actual positives. It is calculated as the ratio of true positives to the sum of true positives and false negatives.

$$\text{Recall} = TP / (TP + FN). \tag{3}$$

Finally, the F1-score provides a balance between precision and recall and is calculated as the harmonic mean of precision and recall.

$$\text{F1-score} = 2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall}). \tag{4}$$

The F1-score is a single metric combining precision and recall, offering a balanced evaluation of the model's performance. It is advantageous in scenarios where we want to consider both precision (minimizing false positives) and recall (minimizing false negatives) as essential factors in the model's effectiveness.

These evaluation metrics provide insights into different aspects of model performance, allowing for a comprehensive assessment of the trained models' effectiveness.

4.2 | Classification Algorithms

Here, three different algorithms were selected for comparison: logistic regression, SVM, and ANN. These algorithms were chosen based on their suitability for classification tasks and previous disease detection applications.

The logistic regression model was trained using the processed dataset, and its performance was evaluated using a learning curve. The learning curve, as depicted in *Fig. 3*, illustrates the relationship between the model's performance and the number of training examples. Initially, with a small number of training examples, the model may exhibit high training error and low validation error, indicating overfitting. However, as training examples increase, the learning curve converges the training and validation error rates. This convergence suggests that the model learns from the data and generalizes well, indicating a balanced and effective performance. The learning curve analysis provides valuable insights into the logistic regression model's behavior. It helps assess the impact of data size on its performance, aiding in decision-making for further model improvement and data collection strategies.

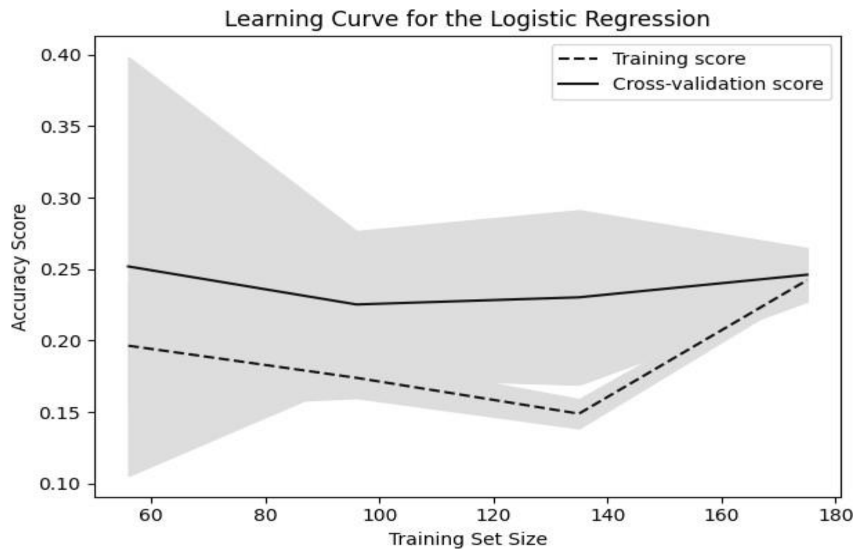

Fig. 3. Learning curve depicting the performance of logistic regression.

A learning curve depicting the SVM's performance was also generated. The learning curve, as shown in *Fig. 4*, provides insights into how the SVM model's performance evolves with varying amounts of training data. With its classification process, the SVM model was specifically designed to distinguish between individuals with Parkinson's disease and healthy individuals within the dataset. During the training process, the SVM model's hyperparameters, including the kernel type and regularization parameter choice, were fine-tuned to achieve optimal performance. The learning curve analysis in *Fig. 4* allows for assessing the ability of the SVM model to generalize from the data. It provides valuable information on its performance characteristics, aiding in the evaluation and refinement of the model.

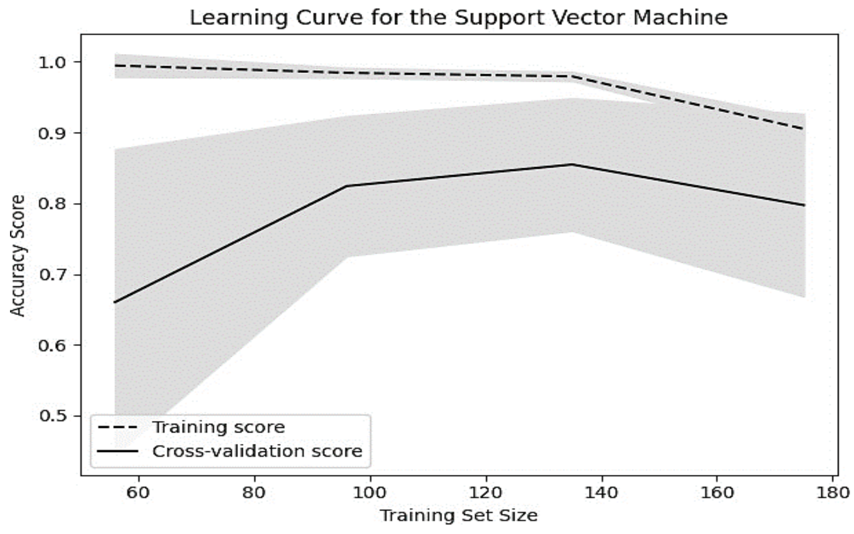

Fig. 4. Learning curve depicting the performance of SVM.

Furthermore, a learning curve was generated to depict the performance of the ANN model. This learning curve, illustrated in *Fig. 5*, provides insights into how the ANN model's performance evolves as the amount of training data varies. The neural network's architecture consisted of multiple layers of interconnected neurons, allowing it to capture and learn complex patterns and relationships within the dataset. The model learned from the data through an iterative process known as backpropagation, where the weights and biases of the neural network were adjusted to minimize the error between the predicted outputs and the true labels. The learning curve analysis offers valuable information on the ANN model's ability to learn and generalize from the data, facilitating the assessment and optimization of its performance.

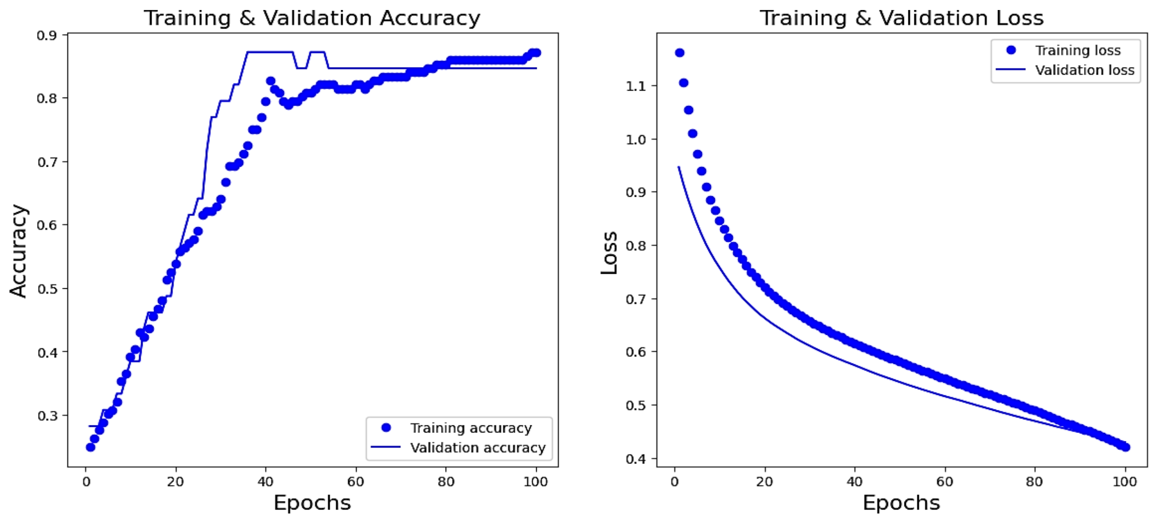

Fig. 5. Learning curve depicting the performance of ANNs.

The following table and figure provide a comparative analysis of the accuracy achieved by different algorithms.

Table 1. Comparative analysis of logistic regression, SVM, and ANN algorithms for Parkinson's disease detection based on accuracy.

Algorithm	Accuracy
Logistic Regression	76.3
SVM	82.06
ANN	92.4

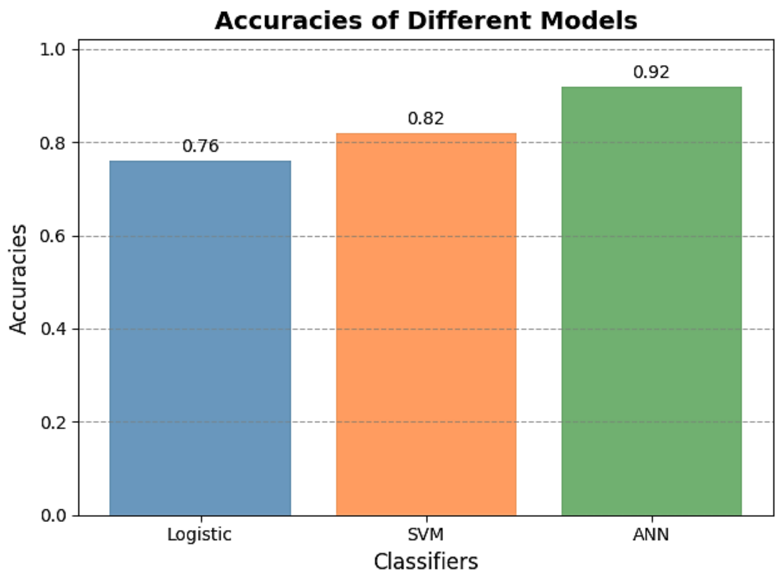

Fig. 6. Comparative analysis between three models.

5 | Conclusion

In a nutshell, this paper provides a roadmap for Parkinson's disease detection using machine learning algorithms. Through a comparative analysis of logistic regression, SVM, and ANN algorithms, we have identified the ANN approach as the most accurate, achieving a remarkable accuracy of 92.4. This study establishes the reliability and efficiency of machine learning models, particularly the ANN, in quickly detecting

Parkinson's disease by analyzing complex patterns and features in patient data. Moving forward, this research lays the groundwork for future investigations in Parkinson's disease detection. Further exploration of machine learning algorithms can be pursued to explore alternative or improved approaches to the ANN model. Algorithms such as random forests or gradient boosting could be evaluated for enhanced performance or interpretability [14]. Additionally, other diagnostic techniques and biomarkers should be integrated to augment the accuracy and reliability of Parkinson's disease detection. Incorporating imaging data, genetic markers, or wearable sensor data can provide valuable insights and improve diagnostic capabilities.

Moreover, it is worth emphasizing the importance of utilizing deep learning methods, such as ANNs, in Parkinson's disease detection. Deep learning models excel at automatically extracting intricate patterns and representations from large and complex datasets. In the case of Parkinson's disease, where the underlying mechanisms and biomarkers can be multifaceted, deep learning approaches offer significant advantages in capturing subtle relationships and achieving high predictive accuracy. Therefore, Future work should focus on leveraging the power of deep learning architectures, such as Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs), to further advance the field of Parkinson's disease detection.

To ensure the robustness and applicability of the models, future work should also focus on validating them on independent datasets from diverse populations. It will enable the evaluation of generalizability across different patient cohorts. Efforts should also be directed towards developing user-friendly and accessible machine-learning tools for healthcare professionals [15]. User-friendly interfaces and decision support systems can facilitate the implementation of these algorithms in clinical settings, enabling efficient and accurate diagnosis of Parkinson's disease. In conclusion, the findings presented in this paper provide a foundation for future research in Parkinson's disease detection. By refining machine learning algorithms, integrating additional data sources, and validating the models on diverse populations, including deep learning methods, we can continue advancing the field and improving early detection and diagnosis of Parkinson's disease, ultimately leading to better patient outcomes and management.

References

[1] Belić, M., Bobić, V., Badža, M., Šolaja, N., Đurić-Jovičić, M., & Kostić, V. S. (2019). Artificial intelligence for assisting diagnostics and assessment of Parkinson's disease—a review. *Clinical neurology and neurosurgery*, *184* 105442. https://www.sciencedirect.com/science/article/pii/S0303846719302380

[2] Schwalbe, N., & Wahl, B. (2020). Artificial intelligence and the future of global health. *The Lancet*, *395*(10236), 1579–1586.

[3] Lalmuanawma, S., Hussain, J., & Chhakchhuak, L. (2020). Applications of machine learning and artificial intelligence for COVID-19 (SARS-CoV-2) pandemic: a review. *Chaos, solitons & fractals*, *139*, 110059. https://www.sciencedirect.com/science/article/pii/S0960077920304562

[4] Saravanan, S., Ramkumar, K., Adalarasu, K., Sivanandam, V., Kumar, S. R., Stalin, S., & Amirtharajan, R. (2022). A systematic review of artificial intelligence (AI) based approaches for the diagnosis of Parkinson's disease. *Archives of computational methods in engineering*, *29*(6), 3639–3653. https://doi.org/10.1007/s11831-022-09710-1

[5] Maulud, D., & Abdulazeez, A. M. (2020). A review on linear regression comprehensive in machine learning. *Journal of applied science and technology trends*, *1*(2), 140–147.

[6] Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W. (2018). Applications of support vector machine (SVM) learning in cancer genomics. *Cancer genomics - proteomics*, *15*(1), 41–51. http://cgp.iiarjournals.org/content/15/1/41.abstract

[7] Jung, S. K., & Kim, T. W. (2016). A new approach for the diagnosis of extractions with neural network machine learning. *American journal of orthodontics and dentofacial orthopedics*, *149*(1), 127–133. https://www.sciencedirect.com/science/article/pii/S0889540615011683

[8] Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC medical informatics and decision making*, *19*(1), 281. https://doi.org/10.1186/s12911-019-1004-8

[9] Mei, J., Desrosiers, C., & Frasnelli, J. (2021). Machine learning for the diagnosis of Parkinson's disease: a review of literature. *Frontiers in aging neuroscience*, *13*. DOI:10.3389/fnagi.2021.633752

[10] Sayed, M. A., Cao, D. M., Islam, M. D. T., Tayaba, M., Pavel, M. E. U. I., Mia, M. T., … & Sarkar, M. (2023). Parkinson's disease detection through vocal biomarkers and advanced machine learning algorithms. *Journal of computer science and technology studies*, *5*(4), 142–149.

[11] Landolfi, A., Ricciardi, C., Donisi, L., Cesarelli, G., Troisi, J., Vitale, C., … & Amboni, M. (2021). Machine learning approaches in Parkinson's disease. *Current medicinal chemistry*, *28*(32), 6548–6568.

[12] Costantini, G., Cesarini, V., Di Leo, P., Amato, F., Suppa, A., Asci, F., … & Saggio, G. (2023). Artificial intelligence-based voice assessment of patients with Parkinson's disease off and on treatment: machine vs. deep-learning comparison. *Sensors*, *23*(4), 1-22. https://www.mdpi.com/1424-8220/23/4/2293

[13] Talitckii, A., Anikina, A., Kovalenko, E., Shcherbak, A., Mayora, O., Zimniakova, O., … & Somov, A. (2021). Defining optimal exercises for efficient detection of Parkinson's disease using machine learning and wearable sensors. *IEEE Transactions on instrumentation and Measurement* (pp. 1–10). IEEE. DOI: 10.1109/TIM.2021.3097857

[14] Magesh, P. R., Myloth, R. D., & Tom, R. J. (2020). An explainable machine learning model for early detection of Parkinson's disease using LIME on DaTSCAN imagery. *Computers in biology and medicine*, *126*, 104041. https://www.sciencedirect.com/science/article/pii/S0010482520303723

[15] Myszczynska, M. A., Ojamies, P. N., Lacoste, A. M. B., Neil, D., Saffari, A., Mead, R., … & Ferraiuolo, L. (2020). Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. *Nature reviews neurology*, *16*(8), 440–456. https://doi.org/10.1038/s41582-020-0377-8